# A General Model for Considering the Variance of CWT Contribution Estimates

Rich Comstock
USFWS

March, 1990

## Introduction

The variance of Coded Wire Tag (CWT) contribution estimates is of central importance to anyone using CWT data. In this paper a model for CWT variance is presented. It is not meant as a competitor to the other estimators wherein bias, efficiency, etc. are considered. Rather it is intended to be used in exploring how the variance of contribution changes under differing assumptions. Using this model assumptions are expressed as covariances between fish or fish groups. This provides a mechanism for considering how behavioral interactions between fish and the dynamics of a fishery can effect the variance of contribution.

## The Population

In CWT studies the populations of interest are often not easy to define. A central issue in describing a population is the notion reproducible results. If we have an estimate of survival for one sample (or group) of fish, selected from a specific population, we would like to think that selection of another, similarly chosen, group from the same population would produce similar results. By similar we mean that the contribution estimates would differ only due to sampling variability. If two simple random samples of all fish in a hatchery were selected for marking, then we would expect contribution estimates computed from these groups to be similar, and we might define the population as all fish in the hatchery at a specific point in time. Usually, however, our intuition of reproducability goes beyond this. Most researchers might like to believe that contributions would be similar if different eggs were selected from the brood stock or if different members of the broodstock managed to escape the fisheries.

Our notion of reproducability often extends over time as well. We often generate contribution estimates for a hatchery stock for several successive years and then use these to predict future survival. The assumption being that all members of a stock have an inherent probability of survival that extends, to some degree, over years.

In the following sections of this paper models of CWT sampling variance are developed. That is, we assume that a population has been defined and that a sample from this population has been selected for marking. We assume also that the population is large in size compared to samples selected for marking and that any marked samples are subject to the same probability distribution as the population. The contribution rate of the population is then estimated by computing a contribution rate of the sample. The sample variance is a measure of the expected difference between the contribution rate computed for the sample and the true population contribution rate. To the extent that a select sample from the population is not representative of (i.e. not subject to the same probability distribution as) the defined population, the results may not be reproducible (i.e. the expected distance between the sample contribution rate and the population contribution rate is greater than that predicted by sampling variance estimators).

CWT studies consist of marking samples from the population and observing the recovery of individuals from these samples. We are interested in a single characteristic of each individual. This characteristic ($Y_i$) is the recovery status of the fish in a specific recovery strata. The recovery strata can be any combination of fisheries, hatchery rack, and/or spawning areas. For fish i, $Y_i = 1$ if fish i was recovered in the recovery strata, and $Y_i = 0$ otherwise. Each $Y_i$ is considered a random variable and $Y = (Y_1, Y_2, \ldots\ldots, Y_R)$ where R is the number of fish in the sample. We assume that each element of the population has the same inherent probability of occurring (i.e. being captured) in the recovery strata. This probability is called the contribution rate. Our goal is to compute a statistic on the vector Y which will estimate this rate.

<u>Notation</u>

R = The number of fish in a marked sample from the population.

a = The number of fish from the sample which occurred (i.e. were captured in) in recovery strata.

p = The probability that a fish from the population occurs in the recovery strata (i.e. the contribution rate).

q = 1 - p

N = Total number of fish which occurred in the recovery strata

n = Number of fish sampled in the recovery strata

s = Number of fish from the sample of interest that were recovered in recovery strata

$$s=\sum_{i=1}^{R} Y_i$$

r = The proportion of fish in the recovery strata that are from the sample of interest

$$r = a/N$$

p, a, r are estimated by:

$$\hat{r}=s/n$$

$$\hat{a}=rN$$

$$\hat{p}=\frac{\hat{a}}{R}=\frac{N}{nR}s$$

Variance of P

For the estimate of p given above, R,N, and n are assumed known. The accuracy of N and n depends upon careful handling of catch and sample records. While some errors will occur it is likely that these will be minimal. Historically, R has often been estimated. The different methods of estimating R preclude the development of a single model of the variance of R. Recently there is a growing trend to use electronic counters to estimate R. If these counters are used carefully, the resulting estimates should be precise enough to allow R to be considered known.

The random variable in the formula for p is s. s is the summation of a sequence of random variables (i.e. $\Sigma Y_i$), each of which has a sample space of 0 or 1. To compute the variance of p we can proceed as follows:

$$\begin{aligned} var(p) &= \frac{N^2}{n^2R^2}var(s) \\ &= \frac{N^2}{n^2R^2}var\left(\sum_{i=1}^{R} Y_i\right) \\ &= \frac{N^2}{n^2R^2}\left[\sum_{i=1}^{R} var(Y_i)+2\sum_{i<j}^{R} cov(Y_i,Y_j)\right] \qquad 1. \end{aligned}$$

$$\begin{aligned} var(Y_i) &= E(Y_i^2)-(EY_i)^2 \\ &= \frac{pn}{N}-\left(\frac{pn}{N}\right)^2 \end{aligned}$$

If we are willing to assume that $cov(Y_i,Y_j)=0$ (i.e. that all pairs of fish are independent) then;

$$var(s) = \frac{Rpn}{N}\left(1 - \frac{pn}{N}\right)$$

and

$$var(p) = \frac{Np}{nR}\left(1 - \frac{pn}{N}\right)$$

Here, var(s) is modeled by the binomial distribution. The binomial distribution is a simple but powerful distribution that describes a great many natural phenomena. The assumptions of the binomial distribution are few. They include: 1) that the characteristic of interest (Y) has two states which can be identified as 0 and 1; 2) the probability that $Y_i$ = 1 is the same for each $Y_i$; and 3) the probability that $Y_i$ = 1 is independent of the value of $Y_j$ (i.e. $cov(Y_i,Y_j)=0$) for all $j \neq i$.

The binomial distribution does not, however, work well for CWT data. The reasons for this can be understood by considering the assumptions of the binomial distribution. First, assumption 2 requires that each element of the population have the same inherent probability of occurring (i.e. being captured) in the recovery strata. This may be referred to as the "single distribution" assumption. Intuitively, there is some reason to question this. Large fish seem to have a different probability of surviving than small fish, etc. It may be feasible to relax the single distribution assumption by considering mixed distribution models. In mixed distribution models different members of the population are assigned different implicit probabilities of occurring in the recovery strata. To date these types of models have been largely unexplored for CWT studies. However, a paper by Newman (1990) may provide a basis for further efforts.

Another reason that the binomial distribution doesn't work well for the CWT is the $Y_i$ are not independent. That is, there are covariances among the $Y_i$s and $cov(Y_i,Y_j)$ <> 0. There are many potential sources of this covariance. By utilizing equation 1, the effects of many of these covariances on the variance of CWT contribution estimates can be modelled. The basic requirement is that we must translate an intuitive source of covariance into a covariance term that operates between pairs of individual fish. Some examples are presented In the following sections.

## Covariance due to Sampling Without Replacement

One source of covariance is due to the sampling without replacement that occurs in fisheries, spawning areas, and hatcheries. Sampling without replacement is modeled by the hypergeometric distribution. The hypergeometric can be derived from the binomial distribution by considering covariances. To accommodate this covariance, the compound binomial-hypergeometric model was developed (Clark and Bernard, Newman 1989). We can use equation 1 to derive the variance of p under these assumptions as follows:

$$E(Y_i)=E(Y_j)=\frac{pn}{N}$$

$$E(Y_i^2)=E(Y_j^2)=\frac{pn}{N}$$

$$E(Y_iY_j)=p^2\frac{n(n-1)}{N(N-1)}$$

$$var(Y_i)=E(Y_i^2)-E(Y_i)^2=\frac{pn}{N}-\left(\frac{pn}{N}\right)^2$$

$$cov(Y_i,Y_j)=E(Y_iY_j)-E(Y_i)E(Y_j)$$

$$cov(Y_i,Y_j)=p^2\frac{n(n-1)}{N(N-1)}-\left[\frac{pn}{N}\right]^2$$

incorporating these results into equation 1 we get;

$$var(p)=\frac{Np}{nR}\left[1-\frac{pn}{N}+(R-1)\,p\,\left[\frac{n-1}{N-1}-\frac{n}{N}\right]\right] \qquad 2.$$

This model is currently used by many CWT data analysts for variance estimation. In the following sections we will use it as a standard for comparison with other models.

Other easily modeled covariance terms arise when different tag codes occur in the same fishery and when one tag group contributes to more than one fishery (Clark and Bernard, Comstock 1989).

<u>Covariance due to Schooling Behavior</u>

The schooling behavior of fish groups can also impose a covariance. For example, it is possible that fish from the same stock, pond, or brood group tend to migrate together throughout their entire life. Hence, if one is taken by a predator it may be that another escapes. If a fish boat captures one, it will likely capture others. The net effect of these circumstances is that the probability of capture of one fish is not independent of that of another. For a simple example of this effect, consider a hypothetical situation where a fish group of size 100 is migrating through a fishery. Suppose that they are migrating in close proximity and that, if the school encounters a fish net, all 100 fish will be caught. Assume that a fish has probability p=0.1 of encountering a net and that all fish in the boat will be sampled for marks (i.e. N=n). If a data analyst were unaware of the schooling behavior of the group then equation 2 may be used to compute variance. Computation of equation 2 yields a variance estimate of var(p) = 0.0009. However, a correct variance is developed as follows;

$$E(Y_i)=E(Y_j)=p$$

$$E(Y_i^2)=E(Y_j^2)=p$$

$$E(Y_iY_j)=p$$

$$var(Y_i)=p-p^2$$

$$cov(Y_i,Y_j)=E(Y_iY_j)-E(Y_i)E(Y_j)=p-p^2$$

incorporating this result into equation 1 we get;

var(p) = pq = 0.09

These variances differ by the inverse of the sample size. The correct variance being the larger.

For a hypothetical example where actual variance less than the binomial model would predict, consider two fish that are migrating together. Suppose they have probability =p of encountering a certain predator. Further, suppose that if this group of two fish does encounter a predator only one fish will be eaten. The variance of the number of fish actually eaten (i.e. $Y.=Y_1+Y_2$) as predicted by the binomial model is 2pq. The actual variance is determined as follows:

$$var(Y.)=var(Y_i)+var(Y_j)+2cov(Y_i,Y_j)$$

$$var(Y_i)=E(Y_i^2)-E(Y_i)^2=0.5p-(0.5p)^2$$

$$cov(Y_i,Y_j)=E(Y_iY_j)-E(Y_i)E(Y_j)$$

$$E(Y_iY_j)=0$$

$$cov(Y_i,Y_j)=-(0.5p)^2$$

$$var(Y.)=2(0.5p-(0.5p)^2)-2(0.05p)^2=pq$$

Here, the actual variance is one-half the variance predicted by the binomial distribution.

## A Simple Model of Fish Schooling and Fishery Dynamics

As a more complex example, we assume that the R fish are evenly divided into m separate schools. Each school has probability $p_0$ of entering the area of a specific fishery. There are T routes through the fishery, N of which are occupied by a fishing net. If a school encounters a net, all fish in the school will be captured. If a boat catches one school then it will cease fishing. Of the N boats in the fishery n will be sampled (note that N and n have been redefined as number of boats rather than number of fish). The variance of p, given these assumptions, is developed as follows:

$$E(Y_i) = p_0\left(\frac{N}{T}\right)\left(\frac{n}{N}\right) = p_0\frac{n}{T}$$

$$E(Y_i^2) = p_0\frac{n}{T}$$

$$var(Y_i) = p_0\frac{n}{T}\left(1 - p_0\frac{n}{T}\right)$$

If $Y_i, Y_j$ are in the same school;

$$E(Y_iY_J) = p_0\frac{n}{T}$$

*and*

$$cov(Y_i, Y_j) = p_0\frac{n}{T} - \left(p_0\frac{n}{T}\right)^2$$

If $Y_i, Y_j$ are in different schools;

$$E(Y_i, Y_j) = p_0^2\frac{n}{T}\left[\frac{n-1}{T-1}\right]$$

*and*

$$cov(Y_i, Y_j) = p_0^2\frac{n}{T}\left[\frac{n-1}{T-1}\right] - \left(p_0\frac{n}{T}\right)^2$$

Of the R(R-1)/2 pairs $(Y_i, Y_j)$, there are m(R/m)(R/m -1) pairs such that $Y_i$ and $Y_j$ are in the same school. There are $m(m-1)(R/m)^2$ pairs such that $Y_i$ and $Y_j$ are in different schools.

Applying these results to equation 1 we get:

$$var(p) = \frac{N^2 p_0}{nmT}\left[1 - p_0\frac{n}{T} + (m-1)\,p_0\ \left[\frac{n-1}{t-1} - \frac{n}{T}\right]\right] \qquad 3.$$

If we set m=100, N=100, n=20, R=10000, T=500, and $p_0$ = .25 equation 3 equals 0.002356. Here p = $p_0$ N/T = 0.05.

Notice that when m=R and T=N equation 3 reduces to equation 2. As a comparison with equation 2 lets assume that there were 1000 total fish captured in the fishery. If the schooling behavior of the fish group and the dynamics of the fishery are ignored then we can use equation 2 with; R = 10000, N=1000, n=200, p=0.05. The resulting variance is 0.00001475. Note that the variances under the two models are different by two orders of magnitude.

The hypothetical situations and models described above are admittedly very simplified approximations to fish migration and fishery dynamics. They are, however, illustrative of vastly more complex covariances that exist in nature. At this time, the effect of naturally occurring covariances within and between fish groups is not well understood. As data is collected and analysis continues, our understanding will improve. The intent of this paper is to argue that equation 1 can be utilized to model the potential effects on var(p) of many types of covariances. However, it is worth noting that the migration and survival of fish groups, as well as propagation of a fish stock, constitutes a complex dynamical system. Being such, it may eventually be found that classical statistical methods will not adequately describe its behavior (Palermo pers. comm.).

## References

Clark, J.E. and Bernard D.R. A Compound Binomial-Hypergeometric Distribution Describing Coded Wire Tag Recovery From Commercial Salmon Catches in Southeastern Alaska. Alaska Department of Fish and Game.

Newman, K. 1989. Variance Estimation for Contribution Estimates Based on Sample Recoveries of Coded Wire Tagged Fish. Proceedings of the Symposium on Marking and Tagging, American Fisheries Society, Bethesda, Maryland (in press).

Comstock, R.M. 1989 Confidence Intervals and Hypothesis Tests for CWT Contribution Estimates Based on the Binomial-Hypergeometric Model. U.S. Fish and Wildlife Service, Olympia, Washington.

Newman, K 1990. Two Competing Models for Hatchery Fish Contribution Rates: A Simple Bernoulli and a Mixture of Bernoullis. U.S. Fish and Wildlife Service, Olympia, Washington.